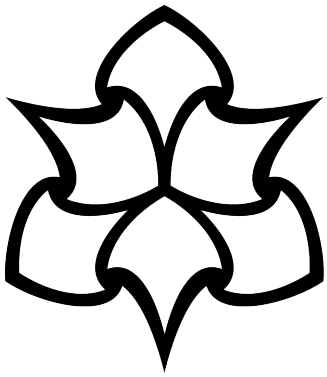


2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)
Bari, Italy. October 6-9, 2019

An Automated Text Mining Approach for Classifying Mental-Ill Health Incidents from Police Incident Logs for Data-Driven Intelligence

Muhammad Salman Haleem[1], Liangxiu Han[1,2], Peter J. Harding[2] and Mark Ellison[1]

Abstract—Data-driven intelligence can play a pivotal role in enhancing the effectiveness and efficiency of police service provision. Despite of police organizations being a rich source of qualitative data (present in less formally structured formats, such as the text logs), little work has been done in automating steps to allow this data to feed into intelligence-led policing tasks, such as demand analysis/prediction. This paper examines the use of police incident logs to better estimate the demand of officers across all incidents, with particular respect to the cases where mental-ill health played a primary part. Persons suffering from mental-ill health are significantly more likely to come into contact with the police, but statistics relating to how much actual police time is spent dealing with this type of incident are highly variable and often subjective. We present a novel deep learning based text mining approach, which allows accurate extraction of mental-ill health related incidents from police incident logs. The data gained from these automated analyses can enable both strategic and operational planning within police forces, allowing policy makers to develop long term strategies to tackle this issue, and to better plan for daytoday demand on services. The proposed model has demonstrated the cross-validated classification accuracy of 89.5% on the real dataset.

I. Introduction

Automatic data-driven intelligence can play a pivotal role in enhancing the effectiveness and efficiency of policing [8]. Police organizations are a rich source of data which varies from crime reports and offender/victim records to more narrative based qualitative data such the records of communication relating to each *incident* (defined as any call for police service). These records can be in structured format (e.g. spatial/temporal information, crime type) as well as in unstructured format (e.g. communication logs between officers on the scene and radio operators/call handlers) [1]. Here we examine incident text logs, which represent a periodic timeline of interactions between officers, *nominals* (e.g. victim, perpetrator, witness, informant) and partner agencies reported back through radio operators and recorded as the incident unfolds. These records are entirely comprised of 'second hand information', in that person(s) recording the data are not those on the scene. They are unique with respect to most text mining datasets in that they are recorded in sequential timestamped blocks, but are not edited after the fact, therefore any information recorded at the start of the log may wholly contradict that recorded at the end. Incident logs are also of varying temporal length, in that one incident may occur over the case of just minutes, whereas others may run the course of days before they are closed. In addition to these issues, the nature of police work means that incident logs are also compiled by 'experts' that are working to exceptionally strict timescales, therefore the presence of abbreviations, organizational acronyms, and non-standard terminology are more commonplace than in standard datasets. These logs are compiled and held, in a similar way, across most regional and international police forces. However, these datasets may have nuances resulting from historical, procedural, organizational, and technical differences between jurisdictions. For example, call handlers can assign 'closing codes' which not only describe the nature of the incident (e.g. firearms present, domestic abuse) but also they are needed to trigger events in which specialist officers will investigate the scene regardless of whether a formal crime has been logged. These codes are therefore not flawless, and it cannot be certain that flagging an incident as being *mental-ill health* (MH) or *non mental-ill health* (NMH) related is far from guaranteed. Therefore the actual time spent by the police officers dealing with MH related incidents are often variable and highly subjective [6]. Since persons suffering from mental disorders are statistically more likely to be in contact with the police officers, both as perpetrators or victims of crime [5], the development of automated text mining approaches to data extraction has the potential to improve operational efficiency and effectiveness.

In this paper, we present an automated text mining approach, based on deep learning, that can identify incidents related to MH. Firstly, we have developed a training set of manually classified incidents (classified as either MH or NMH) by both expert (police) and non-expert (civilian) users and finalize the classification on mutual consensus between operators. Secondly, we have demonstrated how to best represent each word according to its context, and investigate the best method for word embeddings which represent each word and text log in a contextual vector alongside all other words present in each log. Thirdly, we describe the development the convolutional neural network to classify the text logs according as either containing MH component or not, using the pre-defined word embeddings as its input data.

The paper is organized as follows: Section II describes the previous work relevant to classifying this type of text. Section III defines the methodology used to design, develop, train, and test the system for MH classification of police incident logs.

[1] Crime and Well-Being Big Data Centre, Manchester Metropolitan University, Manchester, UK m.haleem@mmu.ac.uk [2]School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University, Manchester, UK l.han@mmu.ac.uk
[2] Elements Technology Platforms Ltd, Sheffield, UK

Section IV presents a summary of how the system performed in its classification tasks on our data-set.
Section V presents our conclusions, and discusses the best practice for this kind of classification tasks, demonstrated empirically throughout the course of this work.

II. Relevant Works

Most of text mining methodologies for identifying MH conditions from text have been developed for social media (twitter, facebook etc.), clinical texts or online forums [7]. Latent Dirichlet Allocation (LDA) [4] has been one of the most common techniques for topic analysis, e.g. person's feeling and intention through online posts [25], depression detection from Twitter activity [26] and social anxiety from therapeutic emails [13]. He et al. [12] used a keywordbased approach to determine if the authors' of self-narratives written by trauma survivors were suffering from post traumatic stress disorders (PTSD). Park et al. [22] examined topic similarity to identify anxiety, depression and PTSD from Reddit posts while correlating different sub-topics within these mental disorders to determine co-occurrence of common symptoms. Larsen et al. [18] took a different approach and used Principal Component Analysis (PCA) to examine correlations between the temporal fluctuations of emotion across twitter and MH related keywords. The majority of these topic modelling applications have been applied on short texts from social media, but we do not believe they are directly applicable to police text logs which possess significantly longer word counts and where a single log can contain mixed information on multiple topics.

Despite of text mining being identified as having a potential towards crime and incident data analysis [8], there are relatively few text mining techniques developed for analysing descriptive police data. Traditionally, historical police reports have been analysed to identify hidden crime patterns [2]. Recently, a keyword annotation based tool has been used to extract MH related cases from online reports of domestic abuse incidents. For this, a behavioural disorders dictionary was required which was compiled for the study by psychiatrists and clinicians [16]. However, dealing with the negation of MH concept was questionable. In other work, supervised text classification of MH was been performed on clinical texts [14]. In both cases, the dictionary of the terms present in the clinical text was built by a team of medical professionals or heath-care professionals. A recent study suggests there exists a significance disagreement between psychiatrists and police officers in assessing whether an incident was MH related or not. The majority of the drug-related, or anti-social behaviour incidents are not necessarily MH related, but are more likely to be classified as such by psychiatrists than police officers [23].

Due to discrepancies in topic modelling and keyword extraction based techniques, and the complexities of building a police, or possibly police force, specific MH dictionary, we instead opt for the more automated approach of *word embeddings* [20]. Word embeddings represent word as a *A*-dimensional vector which quantifies its semantic context within the documents. In the case of our data, word embeddings represent police terminology, abbreviations, codes, and acronyms as a group of multi-dimensional vectors within the context of their surrounding text, negating the need for explicitly classified lists of keywords. We use the word embeddings defined from our training set as input to a

Convolutional Neural Network (CNN), that will be trained to perform the final classification of each log as either MH or NMH based on the criteria it learned during the supervised training phase. The advantage of CNN for this task is their inherent tendency towards multi-scale analysis, which is important for long sections of text that may contain sparsely distributed information from mixed sources. This form of wholly automated, multi-scale approach is, to the best of our knowledge, entirely novel, and in contrast to the previous techniques which focus on dictionary based analysis, word occurrence and frequencies [3], [11].

III. Methodology

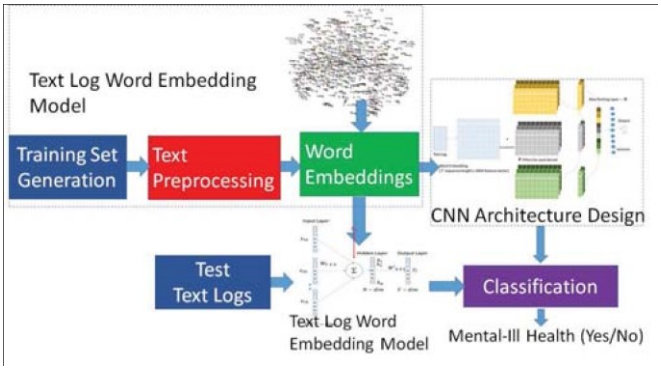


Fig. 1: Block Diagram of the Methodology

The workflow for developing this automatic classifier is shown in Fig. 1. Firstly, the training set was generated from by manual classification of text logs as either MH or NMH, and all logs were subject to standard text preprocessing techniques. Secondly, all possible word embeddings were compiled from the training set, to represent words and phrases in a more contextual manner. Thirdly, the training set is used to teach a mutli-kernel Convolutional Neural Network (CNN) based classifier to identify MH related text logs based on word embeddings. The details are as follows:

A. Training Set Generation and Text Preprocessing

The police text logs were reviewed by multiple persons, and classified as either MH or NMH. The agreement between reviewers assessed using Cohen's κ coefficient, a measure of inter-rater agreement for categorical data [9]. The κ coefficient can be given as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (1)$$

where p_o is observed proportionate agreement, i.e. the proportion of data items for which the reviewers' agreed, whereas p_e represents the agreement that may have been expected if classifications were random. After the incident logs are classified, they are cleaned and standardized by means of

text normalization, in this case was in the form of lemmatization followed by text tokenization [17].

B. Text Log Word Embeddings

We have seen that an incident log can be represented as a sequence of words, that could each be assigned a unique numerical value. These numerical values represent each possible word within the corpus, but each word relies not only on its dictionary meaning, but also on the context of its usage. To retain the context in which each word occurrence was used, we apply the theory of "word embedding". Each word is vectorised into its *root*, and the associated words found around that occurrence in the text. In our case, we define *Text Log Word Embeddings* as A-dimensional vectors which represent the semantic relationship among the words as well as the text logs in which they are present; representing the word in both local as well as global context of the text log. The text log word embedding is a column of weight matrix D along with the definition of word vector of V unique words i.e. x = $\{x_1,x_2,...,x_k,...,x_V\}$ [19]. The values of text log word vectors are determined by log-linear neural network architecture in which weights of the neural network are iteratively updated according to posterior distribution of the words in the corpus based on backpropagation method. Two most common word architectures are Distributed Bagof-Words (DBoW) and Distributed Memory (DM) [20] (See Fig. 2). In our case, we developed the word embeddings based on DM, whose learning objective is to find word vector representations that are useful for predicting the middle word under a context [20]. The weights matrix W of the neural network has dimension V x N, where N is the hidden layer size. For a given one-hot word vector x_k, the hidden units h will be represented as:

$$h_k = D_k + \frac{1}{C}\mathbf{W}^T(x_{1k} + x_{2k} + ... + x_{Ck}) = D_k + \frac{1}{C}(\mathbf{v}_{w_{1k}} + \mathbf{v}_{w_{2k}} + ... + \mathbf{v}_{w_{Ck}})^T \quad (2)$$

where D_k is text log vector from matrix D, C=total number of words present in the context. For the output layer, we have W of N x 1 dimension. The score of the word w_j at the output layer can be given as:

$$u_j = \mathbf{v}'^T_{w_j} h \quad (3)$$

where v'_{w_j} is the j-th column of matrix W. Finally, the output layer y_j can be represented as posterior distribution of word w_j given context words C:

$$y_j = p(w_j|w_C) = \frac{e^{u_j}}{\sum_{j'=1}^{V} e^{u_{j'}}} \quad (4)$$

In this model, the weights are updated according to minimizing the loss function:

$$E = -\mathbf{v}^T_{w_O}.\mathbf{h} + \log \sum_{j'=1}^{C} e^{u_{j'}} \quad (5)$$

where w_O is output word. On the contrary, the DBoW model has the mirror image of the DM model in which word vectors are generated while maximizing the probability of the context words given the current text log vector.

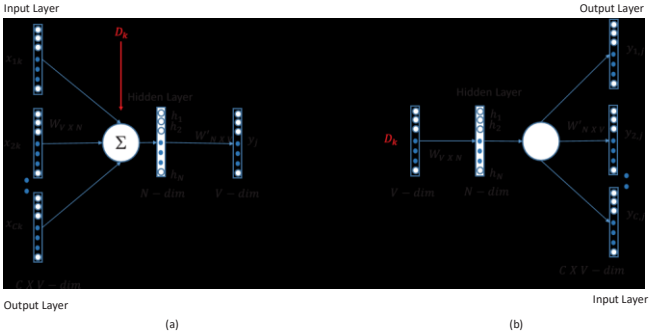


Fig. 2: Pictorial Comparison of (a) Distributed Memory (DM) and (b) Distributed Bag of Words (DBoW) models

The DM based text log word embeddings have been compared with DBoW as well as two most common word embedding models i.e. Continuous Bag-of-Words (CBoW) and Skip-gram models [20]. Both CBoW as well as Skipgram have proved accurate word representations at sentence level classification however, their accuracy at document level classification is questionable [10]. Moreover, they have also been compared with the Global Vectors (GloVe) model which has been developed based on learning co-occurrences of words in the general text [24]. The word embeddings generated by GloVe have proven highly accurate in terms of sentiment analysis and topic analysis in social media (Twitter, Facebook etc.) [15].

C. Multi-Kernel Convolutional Neural Networks

We have developed a multi-kernel Convolutional Neural Network (CNN) architecture for classification between MH and NMH incidents. This model takes text log word embeddings as input for performing text classification according to the word and its context of multiple word lengths while convolving the word sequence with the moving filter. From the output of the word embeddings, we can represent the text logs as concatenation of word vectors in sequence.

$$y_{(1:N)} = y_1 \oplus y_2 \oplus ..... \oplus y_N \quad (6)$$

where $\oplus$ is the concatenation operator. The CNN after convolving with the moving filter generate the feature vector a. For the word vector y_i, we can generate the feature vector a_i while convolving word vectors and its context with the moving filter of size z_k. We have:

$$a^p(z_k) = f(conv(y_{(-\frac{z_k}{2}:\frac{z_k}{2})}, K^p(z_k)) + b^p) \quad (7)$$

where the convolution function can be represented as:

$$conv(y_{(-\frac{z_k}{2}:\frac{z_k}{2})}, K^p(z_k)) = \sum_{r=-\frac{z_k}{2}}^{\frac{z_k}{2}} K_r^p(z_k) y_{i-r} \quad (8)$$

Therefore, eq. 7 after substitution of eq. 8 becomes:

$$a_i^p(z_k) = f\left(\sum_{r=-\frac{z_k}{2}}^{\frac{z_k}{2}} K_r^p(z_k) y_{i-r} + b^p\right) \quad (9)$$

K represent the kernel of moving filter which is initialized with Normal filter, z_k is the kernel size which represent the number of words present in the context and b_p is bias value of the filter p. The full feature matrix A_p is equal to ($K_r^p(z_k)$, can have feature map output of dimension $a^p{}_2(z_k)$, $a^p{}_3(z_k)$,.....,$a^p{}_{N-zk+1}(z_k)$). If we have (N −p Mfilters, we+ 1) x p; where M is number of kernels. In the next layer of the CNN architecture, we can have maximum response of each filter at the maximum pooling layer as follows:

$$o^p(z_k) = f(\max(A^p) + b^p) \quad (10)$$

In case of multiple filters and multiple kernels, we can concatenate the response as follows:

$$O = o^1(z_1) \oplus o^2(z_1) \oplus ..... \oplus o^1(z_2) \oplus ..... \oplus o^p(z_k) \quad (11)$$

The block diagram representation of the mutli-kernel multifilter CNN can be represented in Fig.3.

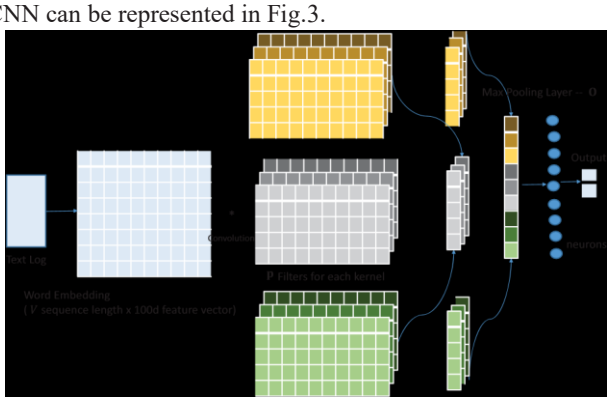


Fig. 3: Multi-kernel Convolutional Neural Networks Architecture

IV. Experimental Evaluation

A. Training Set Generation

The development of the training set involved three reviewers manually classifying logs into MH and NMH categories. Two of the reviewers were from our research team and one from the local police organization. A total of 391 police text logs (average length of 300 and maximum length of 2800 words) have been reviewed. The κ coefficient varies between 0.63 to 0.72 among the reviewers, which demonstrates moderate agreement (see Table I). As a result of the inter-rater agreement, we opted to train based on quorum voting i.e. two of the three reviewers were in agreement. This resulted in 158 MH and 233 NMH examples in the final training set (see Table II). The training set consists of both affirmative and negatory contexts of MH keywords. Affirmative concepts include MH conditions, MH related medications and the suggestions that persons may be at risk suicide or self-harm.

TABLE I: Cohen's κ Coefficient among Incident Text Log Reviewers

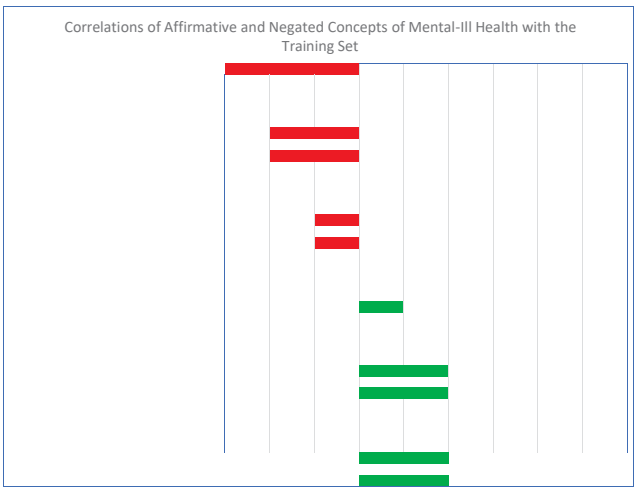
	Cohen's κ Coefficient
Reviewer 1 (researcher) vs Reviewer 2 (researcher)	0.69
Reviewer 2 (researcher) vs Reviewer 3 (police)	0.72
Reviewer 1 (researcher) vs Reviewer 3 (police)	0.63

TABLE II: Text logs labelled as MH and NMH under different categories

Type	NMH	MH	Total
Police Flagged	7	91	98
Police Not Flagged	226	67	293
Grand Total	233	158	391

The correlations between occurrences and classifications of such words can be seen in Fig. 4.

If we compare the results of the manual classifications to those made immediately after the incident by police, we see there is a great difference in classification results. The original "closing codes" suggested that just 93 of the incident logs in our training set had had some form of MH closing code applied, whereas the manual classification after the fact highlighted 153 incidents which had a major MH component. This could be due to many factors, such as the amount of time that reviewers had to make their decisions, the access to the full incident log in a more readable format, or the lack of requirement to apply "actionable" closing codes (codes that trigger further events when attached to incidents). This said, the increase of nearly 61% seen between in the post-facto classification would point to the fact that estimations of the number of police incidents that have a major MH component may be massively under-estimated.



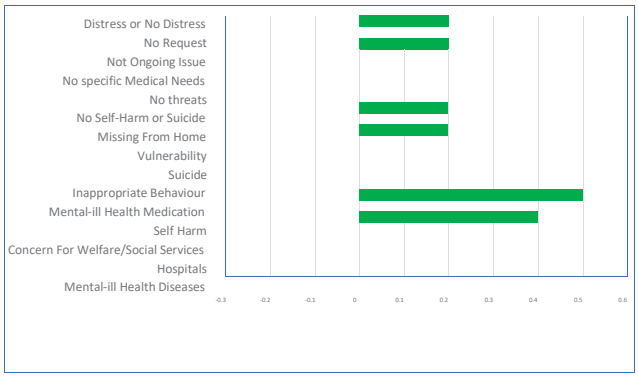


Fig. 4: Correlation Coefficients [21] of Affirmative and Negated MH Concepts with the Training Set

B. Accuracy Evaluation on Different Word Embedding Models

For the comparative analysis, we have performed 5-fold cross validation in which we have divided the training set into

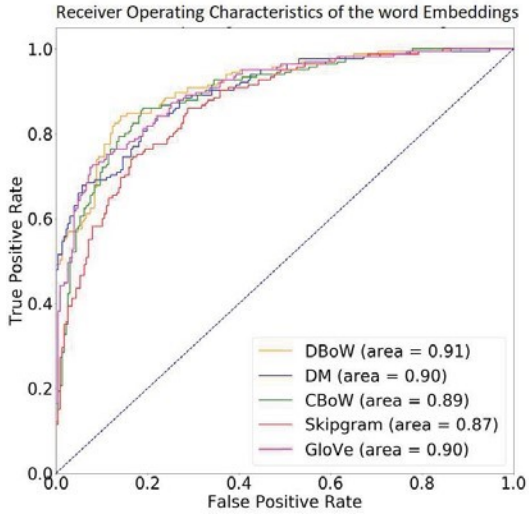


Fig. 5: Comparison of Receiver Operating Characteristics (ROC) for Word Embeddings

5 subsets and each subset has been tested while training the other four subsets. Empirically, we have trained DBoW and DM models while number of words present in the context equals C=30 and generating the 100 dimensional paragraph vectors. We have treated each text log as one paragraph while assigning them paragraph ID. In this way 391 paragraph vectors have been generated. The classification power of text log based word embeddings has been compared (under same parameters of DBoW and DM models) with skip-gram, CBoW [20] and GloVe [24] representing the word at the local context only. Therefore, in order to generate the text logs vectors from these model, we have taken the mean of all the word vectors present in subsequent text logs. These word embeddings have been compared on standard one-layer neural network classifier with number of input and hidden neurons equal to the number of examples present in the training set across each cross validation. The classification power of these word embeddings have been presented by Receiver Operating Characteristics (ROC) in Fig. 5. The results show that document based word embeddings (DBoW and DM) have higher classification power compared to their other counterparts.

C. Accuracy of Word Embedding Models Across Different Kernel Size

We have then determined the classification power across different word embeddings on one-kernel based Convolutional Neural Networks with p=100 filters. These accuracies have also been varied while increasing the kernel size which means increasing number of words in the context. For example, if the kernel size z_k=2, this means the CNN architecture is based on bi-gram k. The results in Fig. 6 provides couple of conclusions: firstly, the generalized Global Vectors (GloVe) may not be applicable for training police text logs due to their different language structure and style. Secondly, CBoW and skip-gram models might be accurate TABLE III: Comparison of Word Embedding accuracies across different multi-kernel CNN architectures

Word Embeddings	z_k=(7,8,9)	z_k=(1-8)	z_k=(2,3,4)
DM	89.5%	84.4%	82.3%
DBoW	85.9%	86.9%	84.9%
CBoW	74.2%	51.4%	73.1%
Skip-gram	83.3%	83.6%	81.8%
GloVe	81.5%	80.03%	81.3%

for sentence level classification however, their accuracy on document based classification and especially on police text logs is questionable. Thirdly, among document based word embeddings, the Distributed Memory (DM) architecture have better classification performance for the higher kernel size whereas the DBoW has better classification accuracy in lower kernel size. As far as highest classification accuracy is concerned, the DM word embeddings managed to have 87.5% on kernel size z_k=8.

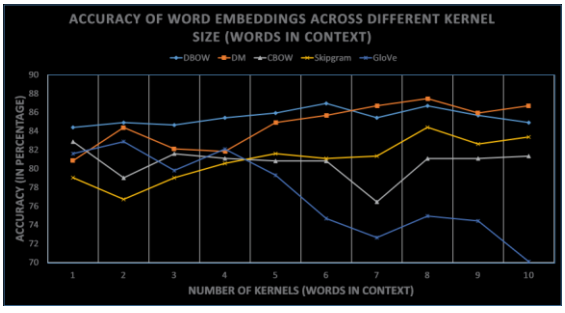


Fig. 6: Comparison of Word Embedding accuracy across different kernel (n-gram) size

D. Multi-kernel CNN Model Evaluation

Due to variable classification performance of DBoW and DM across variable kernel sizes, we have developed the multiple kernel based CNN model. Here we have developed and

compared three architectures. The first architecture was composed of three kernel sizes where *DM* has performed better z_k=(7,8,9). The second architecture was composed of kernel sizes where DBoW has performed better z_k=(2,3,4). The third architecture was composed of all 10 kernel sizes z_k=(1-10). The cross-validation classification accuracies have been presented in the Table.III. The results show that the Distributed Memory based multi-kernel architecture with z_k=(7,8,9) managed to have highest classification accuracy compared to its other counter parts. This shows that document word embeddings based on distributed memory and higher number of words in context can achieve higher classification power in police text logs.

V. Conclusion

We have presented an automatic text mining approach to the classification of mental-ill health related incidents from police text logs. The process by which our training set was developed has been shown, and we have demonstrated (by means of κ) that the classification task is nontrivial, with human operators only being able to agree to a moderate extent. The results of the comparison of postfacto manual classification had increased the number of MH related incident logs by nearly 61%, evidence that the current estimates of MH related demand on the police may be greatly underestimated in the literature.

We have demonstrated that the most commonly used word embedding method (GloVe) is not the best classifier of mental-ill health incidents from incident logs, but that applying a distributed memory model greatly increases classification accuracy. It has also been demonstrated that processing incident logs using both global and local contextual word embedding technique (DM, DBoW) allows for classification with a significantly higher accuracy than processing in just local contextual manner (skip-gram and CBoW). Finally, we have demonstrated that our multiple-kernel based classifier can achieve higher classification accuracy than those which use a single kernel. This final finding is, we believe, a marker of the fact that the incident logs contain data from various sources and relating to a range of subjects. This would lead is to conclude that the multi-scale nature of multiple-kernel based CNN are beneficial to the extraction of data from similar such sources. The final classification has shown that the multiple kernel based CNN architecture achieves a classification accuracy of 89.5% compared to human operators. Future studies include sequential pattern learning based on Recurrent Neural Networks (RNN).

The development of the automatic text mining approaches for the police reports has the potential to reduce time police need to analyse an incident, as well as standardising recording practices. This would allow for evidence-based policy making, intelligence-led policing and for more accurate dayto-day policing decisions such as demand forecasting and budgeting.

VI. Acknowledgement

This research is supported in part by an Economic and Social Research Council (ESRC) grant, reference: ES/P009301/1, and Greater Manchester Police (GMP) who has provided the data.

References

[1] https://data.police.uk/.

[2] S. Ananyan, "Crime pattern analysis through text mining," *AMCIS 2004 Proceedings*, p. 236, 2004.

[3] B. Bengfort, R. Bilbro, and T. Ojeda, *Applied Text Analysis with Python: Enabling Language-aware Data Products with Machine Learning*. " O'Reilly Media, Inc.", 2018.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[5] L. Boulton, M. McManus, L. Metcalfe, D. Brian, and I. Dawson, "Calls for police service: Understanding the demand profile and the uk police response," *The police journal*, vol. 90, no. 1, pp. 70–85, 2017.

[6] K. Bradley, "The bradley report," *Lord Bradley's Review of People with Mental Health Problems or Learning Disabilities in the Criminal Justice System. Department of Health*, 2009.

[7] R. A. Calvo, D. N. Milne, M. S. Hussain, and H. Christensen, "Natural language processing in mental health applications using non-clinical texts," *Natural Language Engineering*, vol. 23, no. 5, pp. 649–685, 2017.

[8] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime data mining: a general framework and some examples," *computer*, vol. 37, no. 4, pp. 50–56, 2004.

[9] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[10] A. M. Dai, C. Olah, and Q. V. Le, "Document embedding with paragraph vectors," *arXiv preprint arXiv:1507.07998*, 2015.

[11] A. Guo and T. Yang, "Research and improvement of feature words weight based on tfidf algorithm," in *2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference*. IEEE, 2016, pp. 415–419.

[12] Q. He, B. P. Veldkamp, and T. de Vries, "Screening for posttraumatic stress disorder using verbal features in self narratives: A text mining approach," *Psychiatry research*, vol. 198, no. 3, pp. 441–447, 2012.

[13] M. Hoogendoorn, T. Berger, A. Schulz, T. Stolz, and P. Szolovits, "Predicting social anxiety treatment outcome based on therapeutic email conversations," *IEEE journal of biomedical and health informatics*, vol. 21, no. 5, pp. 1449–1459, 2017.

[14] R. G. Jackson, R. Patel, N. Jayatilleke, A. Kolliakou, M. Ball, G. Gorrell, A. Roberts, R. J. Dobson, and R. Stewart, "Natural language processing to extract symptoms of severe mental illness from clinical text: the clinical record interactive search comprehensive data extraction (cris-code) project," *BMJ open*, vol. 7, no. 1, p. e012012, 2017.

[15] Z. Jianqiang, G. Xiaolin, and Z. Xuejun, "Deep convolution neural networks for twitter sentiment analysis," *IEEE Access*, vol. 6, pp. 23253–23260, 2018.

[16] G. Karystianis, A. Adily, P. Schofield, L. Knight, C. Galdon, D. Greenberg, L. Jorm, G. Nenadic, and T. Butler, "Automatic extraction of mental health disorders from domestic violence police narratives: text mining study," *Journal of medical internet research*, vol. 20, no. 9, p. e11548, 2018.

[17] T. Korenius, J. Laurikkala, K. Jarvelin, and M. Juhola, "Stemming and¨ lemmatization in the clustering of finnish text documents," in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*. ACM, 2004, pp. 625–633.

[18] M. E. Larsen, T. W. Boonstra, P. J. Batterham, B. O'Dea, C. Paris, and H. Christensen, "We feel: mapping emotion on twitter," *IEEE journal of biomedical and health informatics*, vol. 19, no. 4, pp. 1246–1252, 2015.

[19] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*, 2014, pp. 1188–1196.

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[21] L. Myers and M. J. Sirois, "Spearman correlation coefficients, differences between," *Encyclopedia of statistical sciences*, vol. 12, 2004.

[22] A. Park, M. Conway, and A. T. Chen, "Examining thematic similarity, difference, and membership in three online mental health communities from reddit: a text mining and visualization approach," *Computers in human behavior*, vol. 78, pp. 98–112, 2018.

[23] A. Parker, A. Scantlebury, A. Booth, J. C. MacBryde, W. J. Scott, K. Wright, and C. McDaid, "Interagency collaboration models for people with mental ill health in contact with the police: a systematic scoping review," *BMJ open*, vol. 8, no. 3, p. e019312, 2018.

[24] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
[25] C.-H. Tai, Z.-H. Tan, and Y.-S. Chang, “Systematical approach for detecting the intention and intensity of feelings on social network,” *IEEE journal of biomedical and health informatics*, vol. 20, no. 4, pp. 987–995, 2016.
[26] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki, “Recognizing depression from twitter activity,” in *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 2015, pp. 3187–3196.